Analytical Algorithm To Detect Anomalies With Seasonality Patterns Using Machine Learning Techniques

Sathish Kumar P
PG Scholar
SRMIST, Kattankulathur
Chennai, Tamilnadu, India
sathishkumarbe2013@hotmail.com

Murugan A
Assistant Professor
SRMIST, Kattankulathur
Chennai, Tamilnadu, India
murugan.abap@gmail.com

Abstract—An unexpected observations or items may fall in the dataset. Anomaly detection is the process of identifying dissimilarity between the normal instances. In this paper, it has been analyzed and described an algorithm that achieves a higher and more accurate rate of detection and lowers false positive rates. In recent days, unlabeled data’s are applied as input to anomaly detection algorithms, taking only the internal structures of the datasets. Given a univariate time series, build baselines with lower and upper bounds for the data at a given point in time using historically observed data using zero to minimal input from end users. As the more data are accumulated in time series history, this algorithm finds out the pattern, which automatically determines appropriate lower and upper bounds for the time. The data points which fall outside the lower and upper bounds are flagged as anomalies.

Keywords—univatiate, unlabeled data, anomaly, outlier, bounds

I. INTRODUCTION

In machine learning, “abnormal” instances detection within the dataset provides the consistency of the data. The process of identifying this abnormal instances is said to be outlier detection or anomaly detection. In 1969, Grubbs provides the definition as “An outlying observation, or outlier is one that appears to deviate markedly from the other members of the sample in which it occurs”. Recently with the above definition, the desire to detect that the anomalies may vary with respect to feature. The reason behind the anomaly detection is to eliminate the outliers from the trained data, because pattern reorganization algorithm are sensitive to the outliers in the data, called as data cleaning. As classifiers are deployed in algorithm, the interest to detect outliers were decreased a bit among the detectors. By 2000, the researchers gets more interest in anomaly detection, many optimized algorithm are developed. “In this context, the definition of Grubbs was also extended such that the anomalies should have two important characteristics:

1. Anomalies are different from the normal instances with respect to the feature.
2. Anomalies to be found rare in a dataset compared to the normal instances” [1]

The algorithms to detect the anomalies have been used in various application fields. In Intrusion detection, server applications and the network traffic are monitored to find the anomalies. If the violation occurs in any of the factors, then the anomaly is raised. In fraud detection, log data’s are analyzed to find misuse of the system or suspicious activity resulting fraud. Mostly financial transaction datasets are considered in fraud detection. Data Leakage Prevention is the final obtained application with anomaly detection, where sensitive data’s are protected by detecting data losses.

II. ANOMALY DETECTION SYSTEM
An unexpected observations or items may occur in the dataset. Anomaly detection is the process of identifying dissimilarity for the normal instances. The anomaly detection system with components, involves the concept of dynamic process to detect the data behavior whether the data is normal (defined pattern) or abnormal (anomaly). As described in Fig.1 “Anomaly detection system has three components

- Pre nature components
- Post nature components
- Performance components” [1]

![Figure 1: Anomaly Detection System](image)

Performance of anomaly detection system is accessed by “pre-nature components and post- nature components”. The change in one input affects the performance and the output. Pre nature components calculates truthness of anomalies by using various discrimination criteria’s. Post nature components uses different types of external systems to provide the information about the anomaly instance. Performance components should provide better result in the form of false positive rate, detection rate, accuracy and detection rate. Anomaly detection system determines the expected behaviors depends on type, techniques and application.
A. Pre – Nature Components

Pre - nature components consist of inputs and provides data for analyze, observe and showing road map components for anomaly detection. Some important components are i) Assumptions ii) Statistics iii) Features iv) Algorithm v) Model vi) Activity vii) Constraints viii) Hypothesis.

B. Post – Nature Components

Post - nature components have output parameters of anomaly detection system and provides information towards decision taking. Some important components are i) Decision ii) Alarm iii) Response iv) Analysis v) Feedback

C. Performance Components

Performance components are deriving how effective the anomaly detection system is. Some important considerations of performance are, i) Accuracy ii) Error Rate.

Accuracy

Accuracy is the way of measure to match the evaluated value to the expected one of input applied to the ADS. It provides the response as whether the performance of the systems matches the benchmark level. The system accuracy are based on the assumptions, derived from the models. False Alarm is the important factor for obtaining accuracy. If false Alarm rate is low then the accuracy is better and increases the expected detection rate. By definition “Accuracy of the anomaly detection system depends on the parameters like true positive (TP), false positive (FP), true negative (TN) and false negative (FN)” as shown in Fig. 2 [1]

D. Machine Learning based detection

Machine learning techniques builds a system that considers the performance of previous results to increases the performance of ADS. Owing to the gradually gathered newly information, these systems has the capability to change their strategy of execution. [1] [6]

IV. IMPLEMENTATION

A. Cumulative Sum Control Chart (CUSUM) algorithm

The CUSUM algorithm accurately detects the anomaly using statistical characteristics with smaller computations. This algorithm can detect the change of mean very quickly. [2]

Let $Y_1, Y_2, \ldots, Y_n$ change to identically disseminated free variables over $N(0,1)$ and $Y_{h1}, Y_{h2}, Y_{h3}, \ldots$ be be identically disseminated autonomous variables on $N(0,1)$, where ‘h’ is the obscure change point for the sequence $Y_1, Y_2, \ldots, Y_n$. $\Phi (\cdot)$ demonstrates the dissemination capacity to standard typical dispersion $N(0,1)$. Those probability proportion test the middle of the worth $h=\nu(v<n)$ and the unique suspicion $h=\infty$ will be:

$$
\left\{ \begin{array}{l}
\frac{\prod_{i=1}^{\nu} \Phi(y_i)}{\prod_{i=1}^{\nu} \Phi(y_i)} \\
\frac{\prod_{i=1}^{\nu} \Phi(y_i) \prod_{i=\nu+1}^{n} \Phi(y_i - \delta)}{\prod_{i=\nu+1}^{n} \Phi(y_i)} \\
\end{array} \right. \quad \text{for} \quad h=\nu(v<n) \\
\frac{\prod_{i=\nu+1}^{n} \Phi(y_i - \delta)}{\prod_{i=\nu+1}^{n} \Phi(y_i)} \\
\end{eqnarray}
$$

…equation (1)
\[
V_{n,v} = \exp\left(\delta \sum_{i=v+1}^{n} y_i - \frac{\delta}{2}\right)
\]

Where \(\prod_{i=1}^{v} \Phi(y_i) = 1\) and \(\sum_{i=v+1}^{n} y_i = 0\)

The log value of equation (1) is

\[
\ln V_{n,v} = \delta \sum_{i=v+1}^{n} y_i - \frac{\delta}{2}
\]

... equation (2)

Therefore the log-likelihood proportion the middle of the unique suspicion with no counterbalance and the supposition with counterbalance may be

\[
\ln L_n = \max_{1 \leq c \leq n} A_{n,v}
\]

\[
\ln L_n = \max \left\{ \delta \sum_{i=v+1}^{n} y_i - \frac{\delta}{2} \right\}
\]

... equation (3)

If the sequence is detected with upward shift and \(\delta > 0\), then the ln-likelihood test is equivalent to

\[
Z_n = \max_{1 \leq c \leq n} \left( \sum_{i=v+1}^{n} y_i - \frac{\delta}{2} \right)
\]

... equation (4)

Equation (4) represents the CUSUM value. If the observation values of \(n-1\) are not mean offset, and \(Z_i \leq 0\) where \(i=1,2,...,n-1\) and \(\theta\) is the threshold if \(Y_n, Y_{n+1}, Y_{n+2}, ... Y_{n-1}\) are greater than \(\theta\) is satisfied at time \(n\), then the mean offset is considered to occur.

Let \(\bar{y}_i = y_i - \frac{\delta}{2}\) \(\bar{s}_0 = 0\) and \(\bar{s}_0 = 0\), then the formulae \(Z_n\) can be derived as

\[
Z_n = \max_{1 \leq c \leq n} \left( \sum_{i=0}^{n} \bar{y}_i - \frac{\delta}{2} \right)
\]

... equation (5)

The recursive relation is framed by

\[
Z_n - Z_{n-1} = \left( \bar{s}_n - \bar{s}_{n-1} \right) - \min_{0 < \delta < \theta} \bar{s}_v - \min_{0 < \delta < \theta} \bar{s}_v
\]

... equation (6)

We have

\[
\min \bar{s}_v = \min \left( \bar{s}_v, \min_{0 < \delta < \theta} \bar{s}_v \right)
\]

... equation (7)

Then the recurrence relation becomes,

\[
Z_n - Z_{n-1} = \bar{y}_n - \min_0 \bar{s}_v - \min_{0 < \delta < \theta} \bar{s}_v
\]

... equation (8)

The recurrence formulae for \(Z_n\) is given as

\[
Z_n = \max(0, Z_{n-1} + \bar{y}_n)\]

... equation (9)

For \(\bar{y}_n = y_i - \frac{\delta}{2}\) the uncertainty parameter \(K\) can be replaced by \(\frac{\delta}{2}\), then the recursion \(Z_n\) is derived as

\[
Z_n = \max(0, Z_{n-1} + y_n - k)\]

... equation (10)

If the condition threshold is limited to \(\theta\), and the following conditions are satisfied at the observation point \(h\) \(Z_n > \theta\) where \(Z_i < \theta\) when \(i=1,2,...,h\), then configured output components raises an event or triggers the alarm.

B. Seasonality Pattern Generation Techniques

For detecting an anomaly in the dataset, this algorithm requires at least 30 data points (assuming that every two minutes we hit a data point). By the combination of data points and time sequence, we identify the category of the data. If the accumulated dataset is for more than 2 days, calculate the anomaly based on the daily seasonality. In case, if the accumulated dataset is more than 2 week, algorithm calculates the weekly seasonality. With the seasonality support, algorithm will recognize the pattern very faster, thus reduces the false positive alarms. With the CUSUM algorithm, some rebate is provided, in which it calculates standard deviation, so that slightly deviated data will not be detected as anomaly. The structured way is presented in Fig. 3.

Figure 3: Seasonality based prediction

C. Optimizing Threshold

With the week intensity of traffic, the advance in the apprehension amount in the alpha of arrangement time and optimizing constant \((k, \theta)\) is abundant difficult with the seasonality based CUSUM algorithm. The identified problem
is solved by increasing the threshold with the number of sequence flows. From equation (10), a method will be derived for improving the parameter \((k,\theta)\). The derived method provides much better result with the low traffic intensity. [6][8]

For combination of parameters \((k,\theta)\), if \(y_n - d > \theta\) is satisfied, or if \(y_n + y_{n-1} + y_{n-2} - \sqrt{3d} > \theta\) is satisfied at time \(n\), then the algorithm detects an attack is occurred.

If the new inspection process begins at sequence \(h+1\), then:

\[
Z_n = \sum_{i=h+1}^{n} y_i - \sqrt{n-h}d
\]

...equation (11)

Now the recursion \(Z_n\) is expressed as

\[
Z_n = \max\{0, z_{n-1} + y_n - (\sqrt{n-h} - \sqrt{n-h-1})d\}
\]

...equation (12)

where \(d\) is the constant and ‘\(h\)’ in the training sample is identified by

\[
h = \max\{i: j < n, Z'_j = 0\}\text{and}\ Z'_0 = 0
\]

...equation (13)

V. EXPERIMENTAL RESULTS & ANALYSIS

The unlabeled dataset is considered for validating the anomaly based detection system. The internal structure of the dataset is extracted and the data values with time range for 10 days are fed in the system as the training data. Input training data is shown in Fig. 4

![Figure 4: Training Data Set](image)

The best anomaly detection algorithm in the market fails to hope up with the seasonality data and predicts the much false positive anomalies as shown in Fig. 5

![Figure 5: Anomaly Detection System Output](image)

When the Cumulative sum functionality with the seasonality pattern is enabled, the algorithm doesn’t predict any false positive rate. This algorithm has a glitch in upper and lower bound values because the bounded values are constant for some time as shown in Fig. 6

![Figure 6: With Seasonality Patterns](image)

With the optimized threshold and the seasonality pattern, the algorithm works perfectly by predicting the upper and lower bound values for each and every data in the dataset, thus provides a clear picture about the anomaly in the system.

![Figure 7: With optimized CUSUM + Seasonality Patterns](image)
VI. CONCLUSION

By this paper, our algorithm yields a desirable accuracy of finding false positive in detecting anomalies using seasonality based machine learning technique. This experimental results will have the detection rate accuracy as 98.67% and false positive rate of 1.33% respectively. Setting a very high confidence level can result in making the bands too wide to be useful. Furthermore, this algorithm can be extended with pattern matching and the dataset yielded pattern can be used to detect by early warning.

REFERENCES


